Collecting language data of non-public social media profiles

Jennifer-Carmen Frey, Egon W. Stemle, Aivars Glaznieks
Institute for Specialised Communication and Multilingualism
European Academy of Bozen/Bolzano, Viale Druso 1, Italy
{jennifer.frey,egon.stemle,aivars.glaznieks}@eurac.edu

Abstract

In this paper, we propose an integrated web strategy for mixed sociolinguistic research methodologies in the context of social media corpora. After stating the particular challenges for building corpora of private, non-public computer-mediated communication, we will present our solution to these problems: a Facebook web application for the acquisition of such data and the corresponding meta data. Finally, we will discuss positive and negative implications for this method. [1]

1 Introduction

The exploration of new genres of computer-mediated communication (CMC) has most recently become one of the central research objectives when creating and analysing CMC corpora. Most research projects focus on publicly available language data. For example, there is a lot of research on data such as wikipedia articles and corresponding discussion sites (e.g. Storrer, 2012), public chats (e.g. Beißwenger and Storrer, 2012), twitter statuses (e.g. Greenhow and Gleason, 2012), and public social networking profiles (e.g. Pérez-Sabater, 2012). So far, the attention paid to private conversation in CMC research has been sparse [2], resulting in an under-representation of authentic private communication settings in the current picture of social media language.

The small number of corpora of private CMC may result from various difficulties related to data acquisition. Compared to publicly available data, the acquisition of private data is considerably more difficult in terms of privacy issues[3], technical implementation and sampled data retrieval. Obtaining private CMC data is time-consuming for both the researchers and the participants because direct interaction between the two is needed. Additionally, the data acquisition process may involve various media breaks, this in turn would cause problems in terms of consistency of data transfer and would increase the risk of possible data loss. Consequently, the whole process may turn into a rather expensive endeavour.

However, new forms of data acquisition could help to handle the emerging constraints. Therefore, we developed a method, using technical solutions that rose out of the current settings of media usage, for the acquisition of linguistically relevant social media content. After providing an overview of the underlying research project (Section 2) and listing the most urgent challenges when dealing with individual and user-based data of non-public social media profiles (Section 3), we present our fully integrated web solution, implemented as a Facebook web application (Section 4). Finally, in order to emphasize the relevance of our approach, we discuss its advantages

[2] But see for example the Swiss SMS Corpus http://www.sms4science.uzh.ch

[3] Albeit, thoroughly considering the recommendations on internet research by Markham and Buchanan (2012), for instance, can be exhausting enough.

and disadvantages (Section 5).

2 The DiDi Project

The DiDi project investigates the characteristics of South Tyrolean language use on the Social Networking Service (SNS) Facebook by following a sociolinguistic user-based perspective on language data (Androutsopoulos, 2013). Therefore, the goal is to create a corpus of individual SNS communication that can be linked to other user-based data such as age, web experience and communication habits. We gathered sociodemographic information through an online questionnaire and collected the language data of the entire range of social interactions, i.e. publicly accessible data as well as non-public conversations (status updates and comments with restricted privacy settings, private messages, and chat conversations meaning instant messaging) written and published just for friends or a limited audience.[4] Two month after the release of the app, we ended the data acquisition phase with about 150 users that interacted with the app, offering access to their language data and answering the questionnaire. From those we collected 21.400 private messages, 9.248 status updates (6.784/73% non-public) and 5.399 wall comments (4.622/86% non-public), that matched our specific research criteria (L1 German, living in South Tyrol, texts originated in 2013).

3 Challenges for the Acquisition of non-Public SNS Data for CMC Corpora

Bolander and Locher (2014) and Beißwenger and Storrer (2008) discuss, among others, general issues and challenges for corpora of publicly available CMC data. When dealing with non-public data the stated issues of data acquisition for CMC corpora become more demanding: *legal concerns* add to *ethical issues* already mentioned in previous research, and *technical demands* related to *authentic* data retrieval and the linking of *mixed resources* (i.e. language data and sociolinguistic meta information) get more challenging.

For technical and legal reasons of data acquisition interaction between the user and the researcher becomes an inevitable necessity. Whereas the *legal* situation of the research usage of user-generated language data is still under debate for generally public data, the trend leans towards seeking user consent. User-generated language data is always bound to copyright restrictions therefore making every modification, (re)publication or citation, potentially problematic (cf. Baron et al., 2012). Furthermore, ethical considerations researchers should also respect when doing data acquisition of private personal data, demand that such a consent is to be received in advance and that the user data is anonymised (Beißwenger and Storrer, 2008). For non-public data, this legal and ethical issues are of course even more critical.

But also *technical constraints* make it necessary to interact with the user, to gain access to the data. Most media platforms therefore offer interfaces for third parties to obtain access via an explicit permission from the user. With regard to this, a user consent for the usage of private data is legally – and often technically – necessary.

Finding a *representative sample of participants* for the corpus is another problem that, in fact, many corpus creation projects face. Often expensive public relation campaigns and incentives are necessary to get users to participate in projects where the requested data is personal, often intimate and not written for the public. There are different approaches in gathering the otherwise non-accessible private data, most of them asking for individual submissions of language data by the users as for example in the recent "What's up Switzerland?" project[5]. There, participants of the project need to register and send single threads of conversation via mail, following detailed submission guidelines.

As we wanted to make the participation process as attractive as possible, we tried to find another way to gather the data: Particularly, as we considered this to be tedious for users and researchers, and also troublesome because of privacy doubts on the user side and authenticity doubts on the research side. Speaking of non-public language data, the users might feel that

[4] For a detailed description of the project cf. Glaznieks and Stemle (Submitted).

[5] http://www.whatsup-switzerland.ch/en/

their writing does not reflect "proper" language use, and hence brush it up before donating it. Such modifications however reduce the *authenticity* of the data and should be avoided when analysing the language use in social media.

For the reasons of gaining user consent and sociolinguistic meta-data with the highest privacy for participants (i.e. no personal interaction, no backtracking via mail addresses, etc.) and collecting authentic language data, automatic data collection should be preferred over submission by users. Besides it will make the participation more attractive by simplifying the procedure of sharing language and meta data in an integrated, time-saving and genuine way (i.e. the participation stays within the same platform, using the platform's interfaces and methods that are already familiar for users).

4 Non-Public SNS Data for CMC Corpora – the DiDi Web App

To address the challenges described in section 3, we designed a Facebook web application that manages all the necessary interaction with the participants.[6] A complete run-through consists of the following steps:

1. informing potential participants about the research project, the privacy policy and the data usage declaration;
2. providing options for the user to choose which content to share (private inbox and/or personal wall) and thereby increasing the transparency for the user about which data will actually be retrieved;
3. authenticating the user via the Facebook login dialogue (by using the Facebook API);
4. obtaining the consent to use, save and republish the user's data (via the web application as well as via the Facebook infrastructure for privacy policies);
5. managing the registered user and the granted permissions via the Facebook login dialogue and the Facebook API;
6. requesting an anonymous and individual user identifier for the survey client, saving permission flags, and enlisting the user into an internal database;
7. redirecting to the survey for the acquisition of the user's meta information;
8. providing dynamic feedback to the user about the current progress of the project (e.g. the amount of participants);
9. providing the possibility to share the application with Facebook friends to attract more users.

[6]The source code of the DiDi web application is available at `https://bitbucket.org/commul/didi` for the main application and at `https://bitbucket.org/commul/didi-ws` for the corresponding web service.

5 Properties of an App-Supported Data Acquisition

An app-supported data acquisition has advantageous properties but also some constraints that should be considered.

5.1 Advantages

The most important advantage is that the application facilitates the access to authentic, unrevised and non-public domains of every-day computer-mediated communication. The data is received in a well-defined format and is genuinely machine-readable, easy to restructure or to join with other (social networking) content. Basic annotations, concerning, for instance creation time, privacy settings of content, links to multi-modal elements or devices used for text production, already come with the data.

With respect to the participation process, the web application keeps it as slim and simple as possible. It takes users solely two clicks to donate their language data. After this, the user will be redirected to an integrated online questionnaire. For logging in and accepting the terms of privacy of the app, users do not need to register anywhere but will simply follow the familiar Facebook routines for apps. There is no one-to-one interaction between an authenticated person and a researcher as this would raise privacy issues and doubts in the consistency of anonymisation. Furthermore, legal and ethical constraints are met within the online setting without additional effort. Meta information of the questionnaire and actual language data are automatically linked with an anonymous user identifier, provided by Facebook individually

for every registered user of the app. Therefore, the identifiers can be used even with third-party survey services without privacy problems.

Moreover, the app procedure facilitates the isolation of user acquisition and interaction with the actual crawling of language data. The application only manages registered users. After logging in, the application grants access to the user's account for a period of 60 days. Thus, using such a web application enables efficient data crawling. While users do not have to wait for the language data download to complete, the risk of data loss and other loading and saving issues decreases, as data can be retrieved in independent processes whenever performance and memory capacities allow it best. Furthermore, server or system failures do not result in data loss since the data can be requested repeatedly.

Finally, there are various possibilities to support the attractiveness of the research project. Dynamic feedback can be given through the application surface allowing participants to be part of a collective community project. The application can be easily shared as Facebook post, blog comment, twitter status, e-mail or any other media content. After having finished the survey, participants can directly share the application with their friends via Facebook. This workflow is genuine to social media contexts and addresses interested users wherever they happen to be. In addition, participants can be reached by Facebook via targeted advertising campaigns that address a specific user subset and are usually paid by conversions or actual reach of the advertisement.

5.2 Demands and problems of the application strategy

Using such a web application may save a lot of manual work in data acquisition and be inevitably necessary for the data accessibility. However, it raises the demands on design, development and hosting of the application. Therefore, it increases human workload, required expertise and technical demands. For example, an appropriate infrastructure is needed first of all for the setup of the application (webserver, system and server reliability and monitoring, timely response in case of failures). Secondly, the appropriate infrastructure is needed for a secure and safe data transmission and storage (internal server storage and services, encrypted data transmission and connection, etc.) to ensure anonymity and protect the users' privacy.

In addition to the implementation of the general app functionality and its technical requirements, usability concerns and graphical interface design principles should also be considered to make the software engaging and easy to handle. Therefore, to minimize the efforts in expertise and workload a general app infrastructure for obtaining facebook and/or other social media content as a reusable module for different projects could be a future objective in CMC corpus research.

Another problem within the app approach is the remaining chance of data loss. Within our application design it was not obvious for the users that the data crawling does not happen at the actual moment of participation. The disassociation of these two procedures favours a comfortable participation and crawling procedure, but may also lead to false presumptions. Users may disauthorise the application directly after the participation and hence avert the subsequent data crawling unintentionally. In addition, Facebook is able to refuse data requests even with valid permissions if they suspect the application to be malware. This could occur when downloading a lot of data or when users repeatedly mark the application as untrustworthy. So, there is no guarantee for a complete access to the data during the entire permission period. Thus, the project's ethical and reliable behaviour should be clear and comprehensible.

6 Conclusion

The proposed web app strategy for the acquisition of SNS data facilitates the collection of non-public language data that would otherwise be very complicated or even unfeasible. Therefore, we take our app as a step towards a general and reusable infrastructure that might help to keep the technical efforts for further development low and hence help people to profit from the advantages of this approach.

References

Jannis Androutsopoulos. 2013. Online Data Collection. In Christine Mallinson, Becky Childs, and Gerard Van Herk, editors, *Data Collection in Sociolinguistics: Methods and Applications*, chapter 14, pages 236–250. Routledge, New York.

Alistair Baron, Paul Rayson, Phil Greenwood, James Walkerdine, and Awais Rashid. 2012. Children Online: A survey of child language and CMC corpora. *International Journal of Corpus Linguistics*, 17(4):443–481.

Michael Beißwenger and Angelika Storrer. 2008. Corpora of Computer-Mediated Communication. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook.*, Handbücher zur Sprach- und Kommunikationswissenschaft, chapter 17, pages 292–308. De Gruyter, Berlin.

Michael Beißwenger and Angelika Storrer. 2012. Interaktionsorientertes Schreiben und interaktive Lesespiele in der Chat-Kommunikation. *Zeitschrift für Literaturwissenschaft und Linguistik: Lili*, 42(168):92–125.

Brook Bolander and Miriam A Locher. 2014. Doing sociolinguistic research on computer-mediated data: A review of four methodological issues. *Discourse, Context & Media*, 3:14–26.

Aivars Glaznieks and Egon W. Stemle. Submitted. Challenges of building a CMC corpus for analyzing writer's style by age: The DiDi project. *JLCL*.

Christine Greenhow and Benjamin Gleason. 2012. Twitteracy: Tweeting as a new literacy practice. In *The Educational Forum*, volume 76, pages 464–478. Taylor & Francis.

Annette Markham and Elizabeth Buchanan. 2012. Ethical Decision-Making and Internet Research: Recommendations from the AoIR Ethics Working Committee (Version 2.0). Technical report, AoIR Ethics Working Committee, December.

Carmen Pérez-Sabater. 2012. The linguistics of social networking: A study of writing conventions on facebook. *Linguistik online*, 56(6/12):81–93.

Angelika Storrer. 2012. Neue Text- und Schreibformen im Internet: Das Beispiel Wikipedia. In Juliane Köster and Helmuth Feilke, editors, *Textkompetenzen in der Sekundarstufe II*, pages 277–306. Fillibach, Freiburg.